Big Matters with Small Numbers: Rare Events

Richard Charnigo

University of Kentucky

www.richardcharnigo.net/RE/index.html

Motivating Example

Columns A to E of {DownsSyndromeData.xls} contain information on birth year, maternal age, maternal education, and the presence or absence of Down's Syndrome for 54512 live births in one state from the years 1998 to 2002.

This information was acquired from the National Center for Health Statistics (NCHS) Perinatal Mortality Data Files.

Motivating Example

Over the five-year span there were 59 documented cases of Down's Syndrome among the 54512 live births, 37 among the 17799 live births to mothers aged 30 or older and 22 among the 36713 live births to mothers aged less than 30.

In the year 1998 there were 7 documented cases of Down's Syndrome among the 10782 live births, 4 among the 3499 live births to mothers aged 30 or older and 3 among the 7283 live births to mothers aged less than 30.

Motivating Example

Looking more closely at the 1998 data, the rate of Down's Syndrome was 11.4 per 10000 live births to mothers aged 30 or older and 4.1 per 10000 live births to mothers aged less than 30.

Two questions arise:

- What is the uncertainty associated with these rates?
- Are these rates significantly different?

Motivating Example

Putting aside the possibility that Down's Syndrome may have been underreported, we have to consider how to define uncertainty. After all, the 1998 data are what they are.

One way to define uncertainty is through the following thought experiment. Suppose — contrary to fact — that one more woman aged 30 or older had given birth in 1998. Let p_1 denote the risk of a Down's Syndrome case for that birth.

Likewise, suppose that one more woman aged less than 30 had given birth in 1998. Let p_2 denote the risk of a Down's Syndrome case for that birth.

Motivating Example

We can regard the rates 11.4 (per 10000) and 4.1 as point estimates — single-number "best guesses" — for p_1 and p_2. The uncertainty associated with the 11.4 and 4.1 can then be described in terms of confidence intervals for p_1 and p_2.

In a similar vein, assessing whether the 11.4 and 4.1 are significantly different can entail a test of the null hypothesis that $p_1 = p_2$.

A Failure of Normal-Theory Methods

To construct a 95% confidence interval for p_1 or p_2, we would like to use the familiar and simple normal-theory formula

$$\widehat{p}_1 \pm 1.96\sqrt{\widehat{p}_1(1-\widehat{p}_1)/n_1}$$

or

$$\widehat{p}_2 \pm 1.96\sqrt{\widehat{p}_2(1-\widehat{p}_2)/n_2},$$

where $\widehat{p}_1$ or $\widehat{p}_2$ is the point estimate (expressed as a number between 0 and 1, not 0 and 10000) and n_1 or n_2 is the number of live births.

I have automated this calculation on sheet {Poisson1Risk} of {DownsSyndromeData.xls}. Simply fill in cells G35 through G37. The lower and upper confidence limits then appear in cells J39 and N39.

A Failure of Normal-Theory Methods

We find that the 95% confidence interval for p_1 is 0.2 (per 10000) to 22.6, while that for p_2 is -0.5 to 8.8.

The second confidence interval *looks* wrong, but there is no arithmetic error.

So what is the problem?

A Failure of Normal-Theory Methods

The normal-theory formula assumes that, among n_2 live births with risk of Down's Syndrome p_2 for each live birth, the number of Down's Syndrome cases has an approximately normal distribution.

A mathematical result called the Central Limit Theorem ensures that this is so, but only if $n_2 p_2(1 - p_2)$ is at least 10 — and preferably even larger than that.

We don't know p_2, but we have an estimate available in $\widehat{p}_2$. So,

$$n_2 p_2(1 - p_2) \approx 7283 \frac{3}{7283}\left(1 - \frac{3}{7283}\right) \approx 3.$$

A Failure of Normal-Theory Methods

That we come up with 3 — the observed number of Down's Syndrome cases among live births to mothers aged less than 30 — is no coincidence.

Thus, our operating principle is that the normal-theory formula is only valid if there are at least 10 *cases of Down's Syndrome (or whatever the event of interest is) in the stratum under examination.*

Likewise, the familiar Z test for assessing the null hypothesis that $p_1 = p_2$ is not valid when there are fewer than 10 cases in either stratum.

Poisson Approach to Inference for a Risk

Before we try to perform a hypothesis test involving p_1 and p_2, let us consider performing a hypothesis test just for p_2. Specifically, let us test the null hypothesis that p_2 is 10.0 (per 10000).

If the null hypothesis were true, then the expected number of Down's Syndrome cases among the 7283 live births to mothers aged less than 30 would have been

$$7283\frac{10.0}{10000} = 7.283$$

.

Poisson Approach to Inference for a Risk

Let us assume that, among n_2 live births with risk of Down's Syndrome p_2 for each live birth, the number of Down's Syndrome cases has a Poisson distribution. If the null hypothesis were true, then this Poisson distribution would have mean 7.283.

The assumption of a Poisson distribution is imperfect but decidedly better than the assumption of a normal distribution: quantities that are Poisson-distributed are nonnegative integers, while quantities that are normally distributed need not be nonnegative or integers.

Poisson Approach to Inference for a Risk

Testing the null hypothesis that p_2 is 10.0 (per 10000) then amounts to answering, "How improbable is observing 3 cases when the expected number of cases is 7.283?"

We can answer this question by filling in cells G43 through G45 on sheet {Poisson1Risk} of {DownsSyndromeData.xls} and then examining the contents of cells I148 and J148.

Poisson Approach to Inference for a Risk

Cell I148 informs us that the probability of observing 3 or fewer cases is 0.068 when the expected number of cases is 7.283, while cell J148 reveals that the probability of observing 3 or more cases is 0.976. [Mathematical formulas for these probabilities appear in the Ancillary Notes following this presentation.]

For a two-sided alternative hypothesis, we double the smaller of these numbers — and then round down to 1 if necessary — to obtain a p-value. Here the p-value is 0.136 and we do not reject the null hypothesis that p_2 is 10.0 (per 10000).

Poisson Approach to Inference for a Risk

Exercise: Do we accept or reject the null hypothesis that p_2 is 12.5 (per 10000)? What is the p-value?

Exercise: Do we accept or reject the null hypothesis that p_1 is 12.5 (per 10000)? What is the p-value?

Poisson Approach to Inference for a Risk

Once we know how to test a null hypothesis involving p_2, we can construct a confidence interval for p_2 by the *inversion principle*: any null hypothesis that is accepted at level α is included in the $100(1 - \alpha)\%$ confidence interval, while any null hypothesis that is rejected is not included.

By examining the contents of Column K, we find that a 95% confidence interval for p_2 is 0.9 (per 10000) to 12.0.

Poisson Approach to Inference for a Risk

Exercise: What is a 90% confidence interval for p_2?

Exercise: What is a 95% confidence interval for p_1?

Binomial Approach to Inference for Two Risks

So far we have described how to test a null hypothesis involving p_1 or p_2 as well as how to construct a confidence interval for p_1 or p_2.

What if we want to test a null hypothesis involving both p_1 and p_2, for instance that $p_1 = p_2$?

Binomial Approach to Inference for Two Risks

Fisher's Exact Test is a well-known technique for assessing the null hypothesis that $p_1 = p_2$ when there are very few cases in one or both strata. [A reference for Fisher's Exact Test appears in the Ancillary Notes.]

However, I would like to present an approach that accommodates a greater variety of null hypotheses.

Binomial Approach to Inference for Two Risks

We define the *relative risk* of Down's Syndrome (or whatever the event of interest is) as p_1/p_2 and symbolize it by RR.

The null hypothesis that $p_1 = p_2$ can be expressed as $RR = 1$.

One way to test this null hypothesis is to ask, "How many out of 7 Down's Syndrome cases would be expected to occur to mothers aged 30 or older given that there were 3499 live births to these mothers and 7283 live births to mothers aged less than 30?"

Binomial Approach to Inference for Two Risks

If the null hypothesis were true, the expected number of Down's Syndrome cases occurring to mothers aged 30 or older would have been

$$7\left(\frac{3499}{3499+7283}\right) = 2.272,$$

since the expected fraction of Down's Syndrome cases occurring to such mothers would equal the fraction of such mothers, namely $\left(\frac{3499}{3499+7283}\right)$.

Now we need to answer the question, "How improbable is observing 4 out of 7 cases occurring to mothers aged 30 or older when the expected number for such mothers is 2.272?"

Binomial Approach to Inference for Two Risks

We can answer this question by filling in cells G41 through G43 and cells M42, M43 on sheet {Binomial2Risks} of {DownsSyndromeData.xls} and then examining the contents of cells I1046 and J1046.

Cell I1046 says that 0.960 is the probability that a binomial random variable with parameters 7 and $\left(\frac{3499}{3499+7283}\right) = 0.325$ is less than or equal to 4. Cell J1046 says that 0.160 is the probability that such a binomial random variable is greater than or equal to 4. [Mathematical formulas for these probabilities appear in the Ancillary Notes.]

Binomial Approach to Inference for Two Risks

For a two-sided alternative hypothesis, we double the smaller of these numbers — and then round down to 1 if necessary — to obtain a p-value. Here the p-value is 0.320 and we do not reject the null hypothesis that $RR = 1$.

We can employ the same strategy to test the null hypothesis that $RR = 2.00$, except that the second parameter of the binomial random variable would now be $\left(\frac{2.00 \times 3499}{2.00 \times 3499 + 7283}\right) = 0.490$, since 3499 live births with doubled risk would yield the same expected number of cases as 2.00×3499 live births without doubled risk.

Binomial Approach to Inference for Two Risks

Exercise: Do we accept or reject the null hypothesis that $RR = 2.00$? What is the p-value?

Exercise: Do we accept or reject the null hypothesis that $RR = 20.00$? What is the p-value?

Binomial Approach to Inference for Two Risks

Once we know how to test a null hypothesis involving RR, we can construct a confidence interval for RR by the *inversion principle*: any null hypothesis that is accepted at level α is included in the $100(1 - \alpha)\%$ confidence interval, while any null hypothesis that is rejected is not included.

By examining the contents of Column K, we find that a 95% confidence interval for RR is 0.470 to 18.94.

Binomial Approach to Inference for Two Risks

Exercise: What is a 90% confidence interval for RR?

Exercise: What is a 90% confidence interval for RR if we shift the cut point from 30 to 35 years of age?

Aggregation Approach to Inference

Another strategy that can be used in lieu of — or perhaps in addition to — the Poisson and Binomial approaches previously described is to aggregate or pool data from multiple years (or from multiple geographic regions).

Implicit in this strategy is the belief that p_1 and p_2 are essentially constant across the multiple years (or multiple geographic regions) over which data are aggregated.

Aggregation Approach to Inference

Let us return to sheet {Poisson1Risk} of {DownsSyndromeData.xls} and calculate a 95% confidence interval for p_1 using the data from 1998 through 2002.

Since there are 37 cases among 17799 live births to mothers aged 30 or older, we can employ the normal-theory formula for a 95% confidence interval to find lower and upper confidence limits of 14.1 (per 10000) and 27.5 respectively.

The Poisson approach yields a 95% confidence interval of 14.7 to 28.6.

Aggregation Approach to Inference

Exercise: What is a 95% confidence interval for p_2 using the data from 1998 through 2002, via the normal-theory formula?

Exercise: What is a 95% confidence interval for p_2 using the data from 1998 through 2002, via the Poisson approach?

Aggregation Approach to Inference

Let us continue to sheet {Binomial2Risks} and calculate a 95% confidence interval for RR using the data from 1998 through 2002.

The binomial approach previously described yields a 95% confidence interval of 2.00 to 6.17.

A normal-theory formula,

$$\frac{\widehat{p}_1}{\widehat{p}_2} \exp\left[\pm 1.96 \sqrt{\left(\frac{1-\widehat{p}_1}{\widehat{p}_1 n_1}\right) + \left(\frac{1-\widehat{p}_2}{\widehat{p}_2 n_2}\right)}\right],$$

can also be employed since there are sufficiently many cases in both strata. We obtain 2.047 to 5.878.

Aggregation Approach to Inference

Exercise: What is a 95% confidence interval for RR based on the data from 1998 through 2002 with the binomial approach, if we shift the cut point from 30 to 35 years of age?

Exercise: What if we use the normal-theory formula?

Advantages and Disadvantages of Probability Models

The main advantages of the Poisson and binomial approaches are as follows.

One case in each stratum is sufficient for the validity of all computations described herein, to the extent that we trust the underlying data and are willing to adopt Poisson and binomial probability models for the numbers of cases in a stratum.

Although the Poisson and binomial probability models may be imperfect, they are not "obviously wrong" just because the numbers of cases are small; the same cannot be said of the normal-theory formulas.

Advantages and Disadvantages of Probability Models

Moreover, the Poisson and binomial approaches can be used even when the numbers of cases are not small.

Because one case in each stratum is sufficient, the Poisson and binomial approaches do not require us to assume that p_1 and p_2 are constant over time (or space).

Thus, the Poisson and binomial approaches are useful when we suspect that p_1 and p_2 may be changing over time (or space).

Advantages and Disadvantages of Probability Models

The main disadvantages of the Poisson and binomial approaches are as follows.

Since the numbers of cases may be very small, the point estimates may be wildly unstable from year to year (or region to region).

Although confidence intervals may help us to recognize that some instability in the point estimates is meaningless epidemiologically, they may be so wide as to be noninformative. Likewise, there may be little power for rejecting false null hypotheses.

Advantages and Disadvantages of Aggregation

The main advantages of aggregation follow.

The numbers of cases can often be increased such that the normal-theory formulas may be employed.

Even if one chooses to use Poisson and binomial approaches in conjunction with aggregation, there may be less concern with the point estimates being misleading epidemiologically.

Advantages and Disadvantages of Aggregation

With greater numbers of cases, confidence intervals may be narrowed so that they are informative. Likewise, there may be good power for rejecting false null hypotheses.

Aggregation need not be an all-or-nothing proposition. If one is not comfortable aggregating data over 5 years, one can choose to aggregate data over 3 or even just 2 years.

Advantages and Disadvantages of Aggregation

The main disadvantages of aggregation follow.

The greater the length of time (or region of space) over which one aggregates, the more implausible is the assumption that p_1 and p_2 are constant.

If one's goal is to examine changes over time (or space), aggregation may be antithetical to that goal.

Confidentiality Issues

Some data sets with rare events may, despite not containing traditional identifiers such as names or social security numbers, allow for some individuals to be identified by their neighbors or colleagues based on the combination of a rare event and an unusual mix of demographic characteristics.

This can happen if, for example, there is only one event in a small region occurring to an individual of a minority race in a narrow age stratum.

Confidentiality Issues

The neighbor of such an individual may look at the data set and say, "I know a person of this race and in this age stratum living in this region to whom the event has occurred. So this record in the data set must belong to that person."

Here the problem is not that the event has been revealed — the neighbor already knew of it — but that the neighbor may now be able to look up other characteristics of the person to whom the event occurred, such as that person's income.

Confidentiality Issues

Because of concerns over confidentiality issues, data sets with rare events may be summarized rather than presented in their entireties.

For instance, records for individuals may not be made available, but numbers of events in each of various strata may be furnished along with their respective denominators.

When the denominators themselves are very small, even reporting the numbers of events can be revelatory. In this scenario, the numbers of events — not just records for individuals — may be suppressed as well.

What now?

You're invited to visit the website {www.richardcharnigo.net/RE/index.html}, which contains this presentation and three Excel files with real MCH data. Due to time constraints, we examined only one of those Excel files during this presentation.

You're also invited to review whatever parts of the Ancillary Notes may interest you. They are contained in the same file as this presentation. The Ancillary Notes describe the other two Excel files, present some mathematical details for the Poisson and binomial distributions, and suggest references for further reading.

Ancillary Notes

Besides {DownsSyndromeData.xls}, you can find the Excel files {CleftLipPalateData.xls} and {PostNeonatalData.xls} at {www.richardcharnigo.net/RE/index.html}.

Both of these Excel files contain information acquired from the NCHS Perinatal Mortality Data Files.

Ancillary Notes

Columns A to J of {CleftLipPalateData.xls} contain information on birth year, maternal education, tobacco use, and the presence or absence of cleft lip or palate for 85802 live births in two states from the years 1998 to 2002. I used ten columns because an Excel file is limited to 65536 rows.

Columns A to D of {PostNeonatalData.xls} contain information on birth year, maternal race, and infant mortality for 49857 live births in one state from the years 1997 to 2001. The variable in Column D is coded "0" for infants who survived their first year, "1" for infants who died during their first month, and "2" for infants who survived their first month but died during their first year.

Ancillary Notes

Let μ be a positive number. A random variable X has a Poisson distribution with mean μ if

$$P(X = j) = \exp[-\mu]\mu^j/j!$$

for any nonnegative integer j.

So, for instance,

$$P(X = 0) = \exp[-\mu],$$

$$P(X = 1) = \exp[-\mu]\mu,$$

and

$$P(X = 2) = \exp[-\mu]\mu^2/2.$$

Ancillary Notes

Let us assume that, among n_2 live births with risk of Down's Syndrome p_2 for each live birth, the number of Down's Syndrome cases has a Poisson distribution with mean $n_2 p_2$.

If the null hypothesis $p_2 = p_2^*$ is true, then the expected number of Down's Syndrome cases is $n_2 p_2^*$, which we denote μ^*.

Ancillary Notes

If we actually observe k Down's Syndrome cases, the p-value for the null hypothesis $p_2 = p_2^*$ is given by

$$\min\{1, 2P(X \leq k), 2P(X \geq k)\},$$

where the probabilities are calculated under the supposition that X has a Poisson distribution with mean μ^*.

Explicitly, we have

$$P(X \leq k) = \sum_{j=0}^{k} \exp[-\mu^*](\mu^*)^j/j!$$

and

$$P(X \geq k) = 1 - \sum_{j=0}^{k-1} \exp[-\mu^*](\mu^*)^j/j!,$$

which I have implemented in columns I and J of sheet {Poisson1Risk}.

Ancillary Notes

Let m be a positive integer and q a positive number less than 1. A random variable X has a binomial distribution with parameters m and q if

$$P(X = j) = \frac{m!}{j!(m-j)!} q^j (1-q)^{m-j}$$

for any integer j between 0 and m.

So, for instance,

$$P(X = 0) = (1-q)^m,$$

$$P(X = 1) = mq(1-q)^{m-1},$$

and

$$P(X = m) = q^m.$$

Ancillary Notes

Let us assume that, among m Down's Syndrome cases occurring within two strata, the number occurring within the first stratum has a binomial distribution with parameters m and q.

In this setting q is given by the formula

$$q = \frac{n_1 RR}{n_1 RR + n_2},$$

where n_1 and n_2 are the numbers of live births in the two strata and $RR = p_1/p_2$ is the relative risk of Down's Syndrome.

If the null hypothesis $RR = RR^*$ is true, then

$$q = q^* = \frac{n_1 RR^*}{n_1 RR^* + n_2}.$$

Ancillary Notes

If we actually observe k Down's Syndrome cases in the first stratum, the p-value for the null hypothesis $RR = RR^*$ is given by

$$\min\{1, 2P(X \leq k), 2P(X \geq k)\},$$

where the probabilities are calculated under the supposition that X has a binomial distribution with parameters m and q^*.

Explicitly, we have

$$P(X \leq k) = \sum_{j=0}^{k} \frac{m!}{j!(m-j)!}(q^*)^j(1-q^*)^{m-j}$$

and

$$P(X \geq k) = 1 - \sum_{j=0}^{k-1} \frac{m!}{j!(m-j)!}(q^*)^j(1-q^*)^{m-j},$$

which I have implemented in columns I and J of sheet {Binomial2Risks}.

Ancillary Notes

A good general reference for statistical methods is *Fundamentals of Biostatistics*, Sixth Edition, by Bernard Rosner (Duxbury, 2006).

Section 4.10 introduces Poisson distributions. Equation 14.5 in Section 14.2 presents the Poisson approach to hypothesis testing for a risk but casts it in terms of incidence density for person-time data.

Section 4.8 introduces binomial distributions. Equation 14.9 in Section 14.3 presents the binomial approach to hypothesis testing for equal risks but casts it in terms of incidence densities for person-time data.

Ancillary Notes

Equation 6.23 in Section 6.9 and Equation 6.20 in Section 6.8 demonstrate the inversion principle for the Poisson and binomial approaches respectively.

Section 10.3 presents Fisher's Exact Test to assess a null hypothesis of equal risks.

Equation 6.19 in Section 6.8 and Equation 13.6 in Section 13.3 discuss normal-theory approaches to confidence intervals for risks and relative risks.

Ancillary Notes

Paul A. Beuscher has written a piece called "Problems with Rates Based on Small Numbers" for Issue 12 of the *Statistical Primer* by the State Center for Health Statistics (April 1997). This piece discusses aggregation and indicates that normal-theory formulas will not work when the numbers of cases are too small.

Michael A. Stoto of RAND Health has written a piece called "Statistical Issues in Interactive Web-Based Public Health Data Dissemination Systems" for the National Association of Public Health Statistics and Information Systems (September 2002). Pages 27 through 36 deal specifically with confidentiality issues.